The role of cross-linguistic lexical similarity on bilingual word acquisition

Gonzalo García-Castro[1], Daniela Avila-Varela[1], & Núria Sebastian-Galles[1]

[1] Center for Brain and Cognition, Universitat Pompeu Fabra

Author Note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line.

Enter author note here.

Correspondence concerning this article should be addressed to Gonzalo García-Castro, Ramon Trias Fargas, 25-27, 08005 Barcelona, Spain. E-mail: gonzalo.garciadecastro@upf.edu

Abstract

Previous literature on early vocabulary has been mostly committed to exploring how the receptive and productive vocabulary size changes with age, and its relationship with toddlers' performance on language tasks (e.g. Fernald, Swingley, & Pinto, 2001; Fernald, Perfors, & Marchman, 2006). The content of the developing lexicon has remained relatively unexplored. More recently, studies have focused on the characterisation of the developmental trajectory of individual words, reporting an earlier age of acquisition for words with high frequency, concreteness, and phonological neighbourhood density (e.g. Braginsky, Yurovsky, Marchman, & Frank, 2019; Jones & Brandt, 2019). Such item-level analysis allows not only to predict the age of acquisition of specific words, but also to shed light on the cognitive processes that underlie early word learning. This approach is particularly interesting for investigating the potential impact of acquiring two languages simultaneously. A recent study by Floccia et al. (2018) compared vocabulary sizes of 24-month-old children learning British English, together with an additional language from a pool of 13 diverse languages. They found larger productive vocabulary sizes of toddlers learning two languages that are phonologically similar (see for a similar approach Bosch & Ramon-Casas, 2014). The mechanisms underlying this effect remain unknown. One possibility is that the similarity between the two languages speeds the acquisition of form-similar translation equivalents. The aim of this study is to perform an item-wise analysis on infant's vocabulary contents. We developed an online tool to collect parental reports of receptive and productive vocabularies from children learning Catalan and/or Spanish. We expect that phonological overlap between translation equivalents will predict earlier age of acquisition. If this effect is driven by the phonological overlap between translation equivalents, cognate pairs should be acquired closer in time than non-cognates. For instance, the translation of cat /gato/ in Spanish and /gat/ in Catalan should be learnt at approximately the same age. This should not necessarily happen for the translations of dog, /pero/ and /gos/. We analyse the time elapsed between the acquisition of a word in one language, and its translation in the other. We present preliminary data (data collection is ongoing) on receptive vocabulary of 90 monolingual and bilingual toddlers aged 18 to 36 months. We obtained parental responses to 230 pairs of Catalan-Spanish translation equivalents, resulting in a total of 41490 responses. Words that yielded a phonologically similar translation equivalent were more likely to be present in toddlers' receptive vocabulary at all ages, and were acquired earlier, even when accounting for the effect of word frequency. Further analysis will explore the distance in time between the acquisition of pairs of translation equivalents using a more detailed measure of phonological overlap across translation equivalents and taking into account cross-individual and cross-item variability.

Keywords: word acquisition, vocabulary, bilingualism, lexical similarity

Word count: X

The role of cross-linguistic lexical similarity on bilingual word acquisition

Introduction

One of the main challenges bilingual infants and toddlers face is learning two distinct set of words (one for each language) that partially overlap in sound and meaning. Previous studies have reported that bilingual children know fewer words than their monolingual counterparts when only language is considered (e.g., English monolinguals know more words in English than English-Spanish bilinguals). When taking both languages into account, bilingual children between seem to know, at least, as many words as monolinguals do (**???**; **???**; **???**; **???**; **???**; **???**; **???**). Tough, not all bilinguals show asimilar developmental trajectory of lexical acquisition.

Recent studies have capitalised on the role of the similarity between the specific pair of language the infant is learning. Floccia et al. (2018) analysed vocabulary scores across of 24-month-old toddlers learning English and an additional language. Toddlers learning languages sharing a large amounth of cognates (e.g., English-German) showed larger vocabulary sizes than those learning less languages sharing less cognates (e.g., English-Chinese). (**???**) extended these results to children aged three to 10 years. These results point to the possibility that the overlap between the word inventories infants are acquiring impacts their trajectory of lexical aquisition. However, none of these studies provide an account for the mechanisms involved in this facilitatory effect of language driven by similarity across languages. The aim of this study is to explore two scenarios in which this effect may be taking place:

1) In a first scenario, it is the phonological overlap between language pairs, and not the amount of cognates they share, what boosts the vocabulary growthof bilingual toddlers. Languages sharing a lot of cognates tend to also share a lot of phonemic categories. It is possible that having to learn fewer phonemic categories makes it

easier for toddlers to acquire words. If this is true, toddlers learning two languages with similar phonemic inventories should show larger vocabulary sizes than those learning languages that barely ovelap phonologically. Moreover, this facilitation effect should be reflected in the developmental trajectory of all the items included in a vocabulary inventory, independently of their cognate status.

2) In a second scenario, Words that are form similar It is the amount of cognates language pairs share that boosts vocabulary growth in bilingual toddlers. If this is true, toddlers may acquire cognates earlier

To our knowledge, there are no previous studies that have investigated the effect of cognateness on the acquisition trajectories of individual word forms. Braginsky et al. (2019) used data from MacArthur-Bates CDIs from multiple languages to explore what properties are associated with an earlier acquisition. The found that frequency, concreteness, and [...] significantly predicted an earlier age of acquisition. We plan to extend these analyses to the case of bilinguals.

We have developed a fully-online vocabulary inventory in Catalan and Spanish that is more exhaustive than the MacArthur-Bates CDI, including a larger number of items. All participants filled the questionnaire it both languages, thus providing responses to pairs of translation equivalents.

In this study, we examine the role of cognateness on age of acquisition by modelling the probability of parents reporting that the participant is able to understand, say and understand on each item in the questionnaire. We will test the effect the participants' language profile (monolingual vs. bilingual), and its cognate status on the developmetnal curves of each item.

We will analyse the developmental trajectories of individual words in monolinguals and bilinguals learning Catalan and Spanish (two phonologically close languages sharing a great deal of cognates), between 14 and 30 months of age. We will test whether cognate words are learnt earlier than non-cognate words, taking into account wether they are part of the most- o less-dominant language of the infant, as well as their frequency, concreteness, and number of syllables. We have deloped an on-line questionnaire that entails a short language exposure questionnaire, asks for some demographic information, and presents parent with a list of ~800 words in each language, for which parent report whether their child understands, says, or doesn't understand and say, in each item.

We will fit a Bayesian multilevel model on the item responses, modelling the probability of parents reporting that a given word is acquired by the infant, including language profile (monolingual vs. bilingual), item dominance (dominant language vs. non-dominant language), frequency, concreteness, and number of syllables. We will include *meaning* as grouping variable. We will fit a sigmoidal curve (see **???**) on each item, defined by the steepness of the developmental curve, and the mid-point (the poinnt at which the developmental curve is steepest, which will be considered the point of acquisition). Priors will be derived from previous literature on the subject.

NOTES:

Evidence that bilingual lexica barely overlap (**???**; **???**).

Studies with infants from 8 to 30 months of age show weaker evidence for this claim (**???**; **???**), pointing to the posibility that differences between bilingual and monolingual vocabulary sizes may be dependent on maturational factors.

sugested that the developmental trajectory of lexical aquisition varies across two dimensions: the language profile of the infant, and the joint proporties of the specific language pair the infant is learning. Regarding the first, (**???**) found thant infants exposed to a more balanced

Of special interest is the fact than cognates (form-similar translation equivalents) tend to be overrepresented in the early bilingual vocabulary (Bosch & Ramon-Casas, 2014).

The similarity between both languages may play a role in bilingual lexical aquisition.

Hypotheses

Hypothesis 1 The boost effect of language similarity is driven by cognates being acquired earlier than non-cognates. We predict that age of acquisition of cognates will be ealier thant that of non-cognates. This effect should only be present in bilinguals.

Hypothesis 2 That this effect is driven by the word-forms in one language scaffolding the acquisition of their translation equivalents via parallel activation. We predict that the difference in time between the acquisition pairs of translation equivalents is shorter in cognates than in non-cognates. This effect will only be present in bilinguals.

Method

Participants

Data from 235 participants from the different cities in Spain were invited to participate through social media. Families in the Metropolitan area of Barcelona were also recruited at birth. All families participated voluntarily. Data were collected between 28th October, 2019 and 31th May, 2020, using the BiLexicon inventory. This study was approved by the Comitè d'Ètica de la Investigació amb Medicaments (CEIm) from Hospital del Mar (Barcelona, Spain), code XXXXXXXXX.

- *Location*
- *Age*
- *Sex*
- *Language profile*: : Linguistic profiles were assessed using the Language Exposure Questionnaire (LEQ) (**???**). This questionnaire estimates de degree of exposure (DOE) to each language via a 10 minutes long interview with the parents. LEQs were administered inmediately before or after the testing session at the lab. We here considered only exposure to Catalan and Spanish. Participants with >10% exposure to a third language other than Catalan or Spanish were excluded.

Table 1

Sample demographics and overall language profile. Dominance = Language to which the participant is exposed the most, N = sample size, M age = Mean age, SD age = standard deviation of age, M DOE Cat = mean degree of exposure to Catalan, SD DOE Cat = standard deviation of exposure to Catalan, M DOE Spa = mean degree of exposure to Spanish, SD DOE Spa = standard deviation of exposure to Spanish.

Dominance	Sex	N	M_Age	SD_Age	M DOE Cat	SD_DOE_Cat	M_DOE_Spa	S
Catalan	Female	126,526	20.96	4.59	73.94	18.48	23.52	
Catalan	Male	132,856	20.86	4.78	70.83	17.74	27.77	
Spanish	Female	66,476	21.22	4.76	17.38	14.74	81.14	
Spanish	Male	47,511	20.83	4.88	26.67	13.12	72.66	

- *SES*

Exclusion criteria.

- Language disorders

Questionnaires.

BiLexicon 1.0.

BiLexicon 2.0.

BiLexicon Short. This questionnaire was developed as a short version of the BiLexicon 1.0 and BiLexicon 2.0 inventories. Given the large amount of items to be answered to, we restricted the use of such inventories to participants that were to participate presentially in one of the studies run in the lab at that time. These participants only filled one of the questionnaires, and did so only once.

The aim of developing a short version of this questionnaire (BiLexicon Short) was to provde a more on-line friendly format to families that would participate remotely exclusively, or that would fill the questionnaire more than once (e.g., in longitudinal studies). We divided the items of each language in the pool into four versions named A, B, C, and D. Items in each category (e.g., animals, auxiliary words) were randomly assigned one of the versions. Thos categories that did not entail more than 16 items (thus resulting in less than four items after dividing it in ofur lists), were not divided (all of their items were included in all versions).

When filling the questionnaire for the first time, participants were randomly assigned one version. Each familiy filled only one of the versions, thus reducing the number of items to be answered. To ensure that longitudinal participants would provide longitudinal data in the questionnaire, we made sure that participants were always assigned the same version of this questionnaire. Words that were used as part of the trial lists in the experiment families would participate in were present across all versions.

Item inclusion criteria. We gathered data from 695 distinct translation equivalents across Spanish (695 word forms) and Catalan (703 word forms). The unequal number of word-orms across both languages is due to that some tanslation equivalents did not entail a one-to-one-corespondence: some items had more than one translation equivalent in the other language (e.g. the Catalan word form *walking*, *anar*, can be translated as both *caminar* [to walk] or *ir* [to go]).

We included items from the following categories: Action words, adventures, animals, body parts, clothes, color, descriptive words, food and drink, furniture and rooms, household items, on-line, outside, parts of animals, parts of things, people, pronouns, quantifiers, question words, time, toys, and vehicles. We discarded the following categories: adverbs, auxiliary words, connectives, interjections and games and routines.

We excluded items that enailed more than one word (e.g., *barrita de cereales* [cereal bar]).

Item properties.

Frequencies. We retrieved frequencies from SUBTLEX-ESP (**???**) and SUBTLEX-CAT (**???**) for Spanish and Catalan words, respectively. Word frequency was calculated as the Zipf score of the word (**???**; **???**).

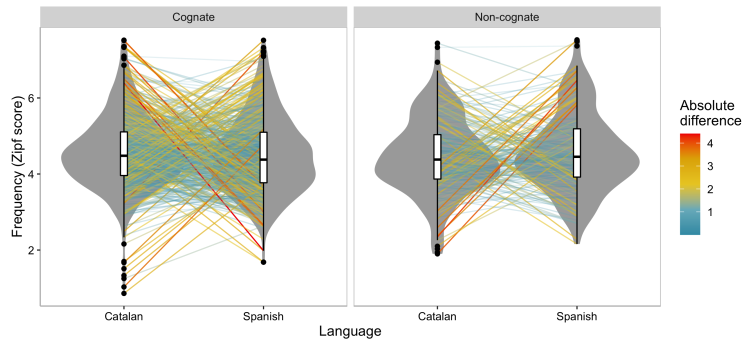

Figure 1

Cognate status.

Demographics by item.

Data analysis

For testing the first hypothesis, we fit a non-linear model using logistic curves to model the probability of aquisition of words across ages. Logistic curves are characterised by three parameters: 1) an asymptote (maximum value of the curve in the Y-axis), 2) the steepness (how fast the curve grows), and 3) a mid-point (the point in the X-axis at which the steepness is highest). Figure 1 shows the elements of a simulated logarithmic curve:

One reason for using logistic curves to fit the data is that previous studies describe lexical development as a non-linear process, where between 20 and 24 months a vocabulary spurt happens. An analysis of the available data in Wordbank (**???**) provides moderate support for this hypothesis. We fitted a logistic model using the proportion of infants that acquired each item as output, and age as input, estimating the three parameters afore mentioned (asymptote, mid-point, and steepness). We included data from 13119 items

Table 2

Demographics of responses by item. Language = item language, LP = language profile of participant, Sex = sex of participant, M Age = mean age of responses, SD Age = mean SD of responses across items, Min Age = minimum age of responses across items, Max Age = maximum age of responses across items, M N = mean number of responses across items, N = total number of responses across items.

Language	LP	Sex	M Age	SD Age	Min Age	Max Age	M N	N
Catalan	Bilingual	Female	22.36	22.36	12.35	31.42	52.22	41620
Catalan	Bilingual	Male	20.77	20.77	12.37	31.91	64.06	51057
Catalan	Monolingual	Female	20.14	20.14	12.83	30.15	68.01	54204
Catalan	Monolingual	Male	21.11	21.11	12.60	30.93	48.30	38495
Spanish	Bilingual	Female	22.33	22.33	12.35	31.42	52.76	42209
Spanish	Bilingual	Male	20.72	20.72	12.37	31.91	64.72	51778
Spanish	Monolingual	Female	20.10	20.10	12.83	30.15	68.71	54969
Spanish	Monolingual	Male	21.06	21.06	12.60	30.93	48.80	39037

from 20 languages[1], from infants between 16 and 30 months of age. Then we compared the fit of the logarithmic model agains a linear model, and found that the formr fitted the data slightly better (see Figure 2).

A second reason for using logistic curves is that the parameters that characterise a logistic curve map onto our hypotheses quite easily. Out first hypothesis concerns the mid-point parameter, which indicates the point at which the steepness of the curve is maximum. In our model, the value of this parameter can be interpreted as the age at

[1] Chinese (Cantonese and Mandarin), Croatian, Czech, Danish, English (Australian, British, and American), French (European), German, Hebrew, Italian, Korean, Latvian, Portuguese, Russian, Slovak, Spanish (European and Mexican), and Turkish. All the other available languages were excluded, as no (or insufficient) data were available.

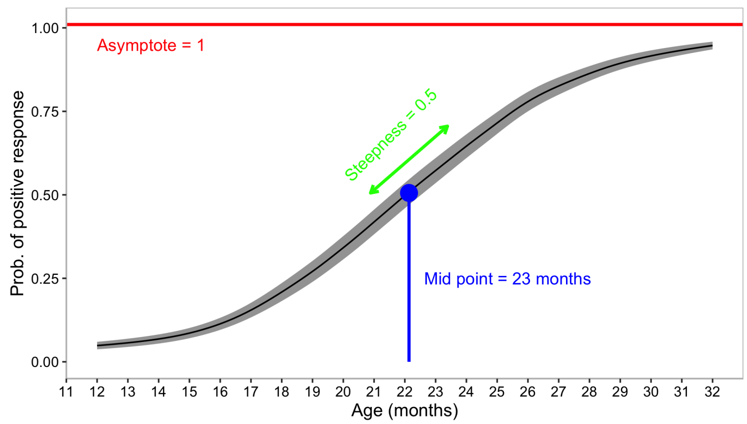

Figure 2. Simulated logarithmic curve. The curve is characterised by three parameters: Asymptote (red), mid-point (blue), and steepness (green). This curve is the result of simulating learning curves of 1000 items. We set fixed values for the asymptote at 1. Mid-points were randomly generated from a normal distribution truncated at 0 (no negative age values are allowed), with mean 23 (*SD* = 0.5), and steepness values were generated from a normal ditribution truncated at 0 (we assume that the proportion of participants that have acquired each item can only grow), with mean 0.1 (*SD* = 1) The solid black line represents the average proportion of participants who have acquired the item, and the shaded ribbon indicates the standard error of the mean. We suggest that learning curves of individual word-forms, as well as the evolution of vocabulary size across ages, follows this trend. Our model will estimate the three of to adress our hypotheses.

which the proportion of participants that have acquired a given item is at maximum rate. The mathematical interpretation of this parameter is easily mapped into a definition of age of acquisition.

We built a Bayesian logistic model that estimates the value of the mid-point and steepness across items. Under our hypothesis, mid-points are generated from a linear model that incorporates language profile (monolingual vs. bilingual) and cognateness (non-cognate vs. cognate) as predictors, where cognateness predictes ealier mid-points for bilingual, but not monolingual participants. Due to computational limitations, we set a

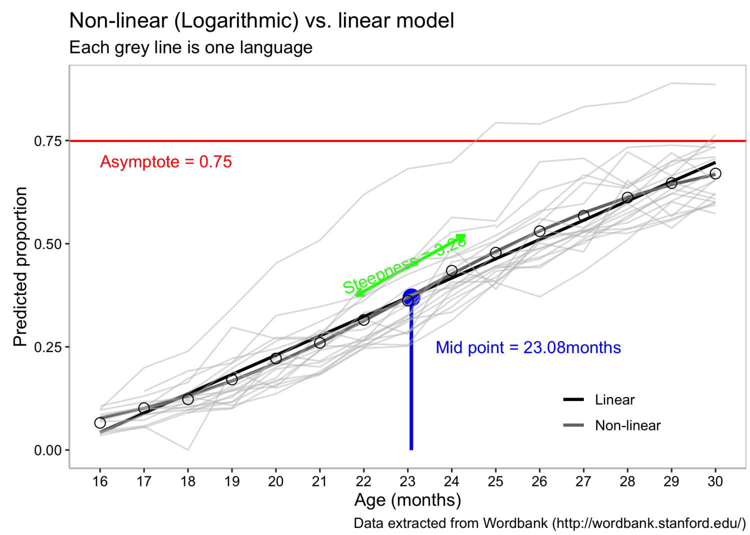

Figure 3. Logarithmic model fitted on Wordbank data, compared to a linear model. The logarithmic model fits the data slightly better. The estimated values of the three parameters involved in the logarithmic model (asymptote, mid-point and steepness) are indicated.

fixed value for the asymptote at 0.75, according to our model fit on Wordbank data. Due to computational limitations, we did not run the same linear model to estimate steepness. Rather, this parameter was just included as an intercept. We compared how well this model fits the data *versus* how a null model that does not include *cognateness* as predictor does, by using Bayes factors and Pareto-smoothed importance sampling (PSIS; **???**).

We used the information provided by Wordbank as priors to feed our Bayesian model. For mid-points, we specified a strong prior with a normal prior with mean = 23.08 and *SD* = 1. For steepness, we also set a strong prior with a normal distribution with mean = 3.28 and SD = 2. We had little prior information about how cognateness impacts the mid-point of an item, and therefore we will set a weak prior.

We fit both null and alternative models within the R environment (**???**; **???**) using the `brms` package (**???**), which relies on the probabilistic language Stan (**???**). Data and model results will be processed and visualised using the R packages the `tidyverse` family of packages (**???**) and the `tidybayes` package (**???**).

Results

Model selection

Posterior distibutions

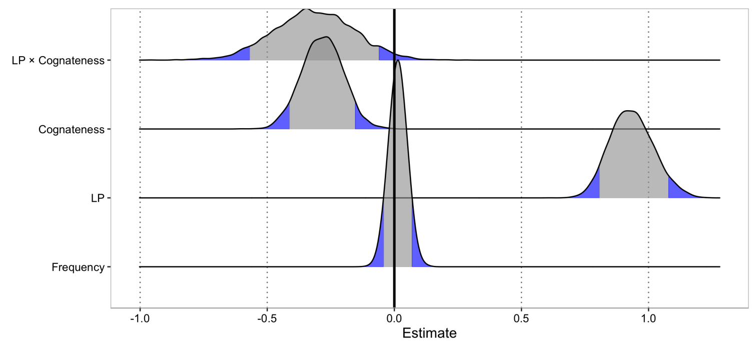

Figure 4

Posterior correlations

Prior predictive checks

Following (**???**), we performed prior predictive chackes to disgnose the appropriateness of our priors.

Posterior predictive checks

Model diagnosis

Auto-correlation.

Potential scale reduction factor ($\hat{R}$).

Posterior predictive checks

Does our model generate similar observations to the ones in our data-set?

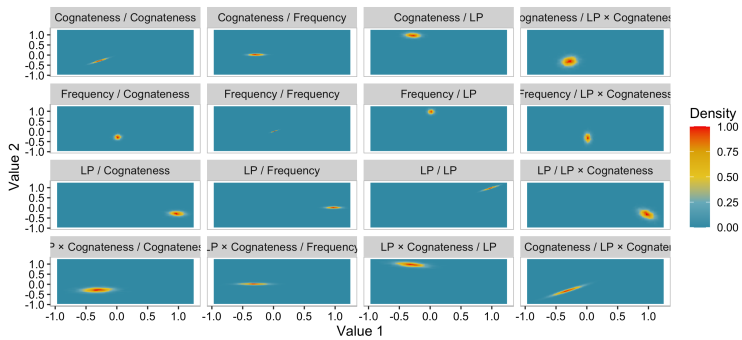

Figure 5

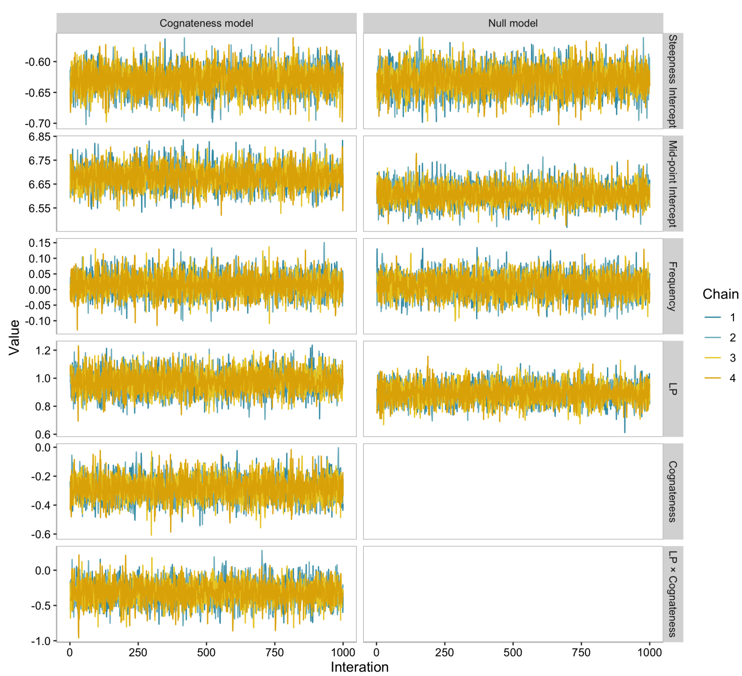

Figure 6

Discussion

Limitations

- Cross-sectional data for longitudinal claims
- Data is right censored
- Inventory:
  - Phonological forms?
  - We can't say how frequently words are produced by the parents or the toddler.
  - Although instructed to ignore imitations, it is difficult to say whether a toddler has *really* acquired a word that produces, or she can only imitate it.
  - Responses in the questionnaire rely heavily on parental memory.
  - We don't know about the context at which words are heard or produced.
  - Classification system of items (e.g., *household items*) is adult centric. Children may use the words *pretty* to name jewellery (Bates et al., 1994).
- Mid and mid-upper class families are overrepresented in the sample.

Appendix

Appendix 1: Session info

R version 3.6.3 (2020-02-29) Platform: x86_64-apple-darwin15.6.0 (64-bit) Running under: macOS Catalina 10.15.4

Matrix products: default BLAS: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib

locale: [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8 attached base packages: [1] stats graphics grDevices utils datasets methods base

other attached packages: [1] rlang_0.4.6 here_0.1 data.table_1.12.8
[4] truncnorm_1.0-8 wesanderson_0.3.6 patchwork_1.0.0.9000 [7] ggridges_0.5.2 ggplot2_3.3.1 lubridate_1.7.8
[10] tidyr_1.0.3 readxl_1.3.1 stringr_1.4.0
[13] forcats_0.5.0 dplyr_1.0.0 tibble_3.0.1
[16] magrittr_1.5 knitr_1.28 papaja_0.1.0.9942

loaded via a namespace (and not attached): [1] Rcpp_1.0.4.6 pillar_1.4.4 compiler_3.6.3 cellranger_1.1.0 [5] plyr_1.8.6 tools_3.6.3 digest_0.6.25 evaluate_0.14
[9] lifecycle_0.2.0 gtable_0.3.0 pkgconfig_2.0.3 yaml_2.2.1
[13] xfun_0.14 withr_2.2.0 generics_0.0.2 vctrs_0.3.0
[17] rprojroot_1.3-2 grid_3.6.3 tidyselect_1.1.0 glue_1.4.1
[21] R6_2.4.1 rmarkdown_2.1 bookdown_0.18 purrr_0.3.4
[25] backports_1.1.7 scales_1.1.1 ellipsis_0.3.1 htmltools_0.4.0 [29] colorspace_1.4-1 stringi_1.4.6 munsell_0.5.0 crayon_1.3.4

Appendix 2: Model

References

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J. S., … Hartung, J. (1994). Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language*, *21*(1), 85–123. https://doi.org/10.1017/S0305000900008680

Bosch, L., & Ramon-Casas, M. (2014). First translation equivalents in bilingual toddlers' expressive vocabulary: Does form similarity matter? *International Journal of Behavioral Development*, *38*(4), 317–322. https://doi.org/10.1177/0165025414532559

Floccia, C., Sambrook, T. D., Luche, C. D., Kwok, R., Goslin, J., White, L., … Plunkett, K. (2018). I: Introduction. *Monographs of the Society for Research in Child Development*, *83*(1), 7–29. https://doi.org/10.1111/mono.12348